BITS 2005: Bioinformatics ITalian Society Conference

Raffaele Ponzini
CILEA, Segrate

Abstract
A Milano dal 17 al 19 Marzo si è svolta la seconda conferenza annuale della BITS (Bioinformatics ITalian Society) [1]. La conferenza, preceduta da una sessione di tutorial sul tema della comparative genomics e sulle prospettive offerte dal Grid Computing rispetto al mondo della bioinformatica, ha avuto in programma diverse sessioni: algoritmi e applicazioni, genomica comparativa e funzionale, bioinformatica strutturale, database e datamining ed infine bioinformatica medica.

The second BITS (Bioinformatics ITalian Society) annual meeting took place in Milan (17 - 19 March) [1]. The conference was preceded by a tutorial session on Comparative Genomics and on the perspectives offered by the Grid Computing with respect to Bioinformatics. The sessions presents were: algorithms and applications, comparative and functional genomics, structural bioinformatics, database and datamining and medical bioinformatics.

Keywords: bioinformatics conference, comparative genomics, medical bioinformatics, structural bioinformatics, functional genomics, grid technology.

Introduzione

La BITS (Bioinformatics ITalian Society) (Fig. 1) [2], è un' associazione scientifica senza scopo di lucro che nasce il 19 giugno 2003 per riunire i ricercatori interessati alla Bioinformatica in Italia. L'associazione si pone come obiettivo primario quello di promuovere lo studio e la diffusione della Bioinformatica sia nell'ambiente scientifico-accademico che in quello tecnologico-industriale.

Figura 1 - Logo della società italiana di bioinformatica

Dare una definizione di Bioinformatica non è semplice in quanto è per sua natura una scienza trasversale, oltre che giovane, che racchiude in sè competenze provenienti da diversi campi teorici ed applicativi. Essenzialmente è una scienza che affronta lo studio dei problemi biologici a livello molecolare e cellulare utilizzando metodi e modelli informatici e computazionali.
In Figura 2 viene data una rappresentazione schematica di questa definizione.

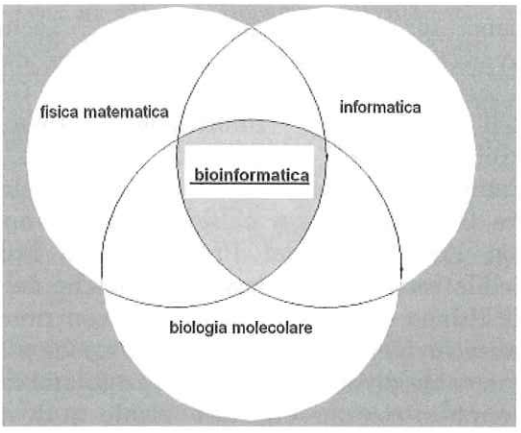

Figura 2 - Schema rappresentativo della Bioinformatica come inter-disciplina.

Nel corso degli anni si è affermata per il contributo innovativo che ha portato alla

biologia moderna sfruttando le potenzialità di affermate tecniche computazionali e numeriche per affrontare problemi di natura biologica. In questo senso la Bioinformatica è diventata un supporto scientifico a discipline biologiche fondamentali quali Biologia Molecolare, Biochimica, Genetica, Biofisica, ecc. e a tutte le discipline di ambito biotecnologico che sempre più suscitano l'interesse del mondo scientifico moderno.

La BITS si propone come punto di riferimento nazionale per il confronto ed il dibattito formativo in questo settore culturale, avendo come fine specifico quello di collocare il nostro paese tra quelli che hanno per tempo riconosciuto il ruolo fondamentale della Bioinformatica ed hanno saputo investire conseguentemente nella formazione della futura classe di studiosi specializzati in questo ambito. Storicamente il Gruppo di Cooperazione Bioinformatica, nato nel 1999 sotto gli auspici della SIBBM (Società Italiana di Biofisica e Biologia Molecolare) [3], prima e l'ABCD (Associazione di Biologia Cellulare e Differenziamento) [4] poi hanno costituito il punto di partenza per la bioinformatica in Italia. L'eredità del Gruppo di Cooperazione è stata ora colta dalla BITS.

La conferenza

Quest'anno la conferenza BITS [1] si è svolta nei giorni 17-19 Marzo. In particolare, come previsto dal programma la conferenza ha avuto inizio il giorno 17 con una sessione introduttiva, ospitata dall'Istituto di Tecnologie Biomediche (all'interno della struttura del LITA, Laboratorio Interdisciplinare Tecnologie Avanzate, Segrate). E' stata aperta da due tutorial: uno tenuto da Mickhail S. Gelfand dell'Institute for Information Trasmission Problems, Ras, Russia, [5] sul tema della Comparative Genomics, l'altro sulla relazione tra Grid Technology e Bioinformatica apertosi con l'intervento del Prof. Luciano Milanesi dell'Istituto di Tecnologie Biomediche del CNR di Milano e seguito da numerosi contributi che hanno evidenziato come esistano anche a livello nazionale diversi progetti in questo ambito, che hanno in comune una forte richiesta di risorse computazionali e di storage di dati.

La sessione si è poi conclusa con un intervento estremamente interessante tenuto da David Lipman direttore del National Center for Biotechnology Information (NCBI) [6], dal titolo "Building Information Resources at NCBI: The Inside Story". L'NCBI, nata nel 1988, è uno storico centro statunitense per la diffusione ed il supporto dell'informazione nell'ambito molecolare, gestisce uno tra i maggiori database del mondo di dati di interesse biologico e molecolare, e, allo stesso tempo, conduce ricerca per lo sviluppo ed il supporto di strumenti di analisi di dati gnomici, permettendo la disseminazione d'informazioni in tutto il mondo biomedico. L'intervento si è svolto in maniera molto informale. Lipman ha saputo catturare l'attenzione dei presenti raccontando aneddoti molto divertenti e stimolanti sull'attività svolta all'interno del suo centro e su come le politiche da seguire nella ricerca dell'informazione richiesta siano essenziali quando si lavora con database delle dimensioni di quello dell'NCBI, confrontando queste politiche con quelle utilizzate dai più importanti motori di ricerca in Internet quali google [7] e la sua versione scholar [8]. Lipman ha anche evidenziato come attualmente il database dell'NCBI non abbia adottato nessuna piattaforma Grid per la fornitura di informazioni e questo è stato motivato con argomentazioni circa gli elevati requisiti di affidabilità e di sicurezza richiesti dal sistema, non ancora raggiunti da queste tecnologie. Secondo lo stesso Lipman inoltre l'interesse negli USA è comunque calante nei confronti della Grid Technology dopo che la nascita delle autostrade informatiche, volute dall'amministrazione Clinton, ne aveva fornito lo spunto allo sviluppo, portando conseguentemente all'entusiastica nascita di progetti fortemente interdisciplinari e di ampio respiro che vedevano questa tecnologia come nodo fondamentale.

Il programma per i giorni successivi, svoltisi presso il centro congressi dell'Hotel Leonardo da Vinci, ha compreso per il giorno 18 due sessioni: una dal titolo Algorithms and Applications, presieduta da Giancarlo Mauri, l'altra dal titolo Comparative and Functional Genomics presieduta da Graziano Pesole rispettivamente dell'Università degli Studi di Milano Bicocca e dell'Università degli Studi di Milano.

Nel giorno 19 infine si sono tenute tre sessioni sui seguenti temi: Structural Bioinformatics, Database and Datamining e Medical Bioinformatics presiedute rispettivamente da Rita Casadio, responsabile del gruppo di Biocomputing dell'Università di Bologna (nodo italiano del progetto europeo

Biosapiens [9]), da Giorgio Valle dell'Università degli Studi di Padova e da Luciano Milanesi.

All'interno di queste due giornate che hanno costituito il cuore della conferenza si sono svolte delle poster session in cui è stato possibile chiedere chiarimenti sui lavori presentati, scambiarsi informazioni e prendere contatti utili.

Interventi selezionati

Tra gli interventi più significativi riguardo all'utilizzo della tecnologia Grid nel mondo della Bioinformatica si vuole commentare quello dal titolo "Biological database access and integration using web services in GRID technology" presentato da Ivan Morelli, Mauro Landenna e Luciano Milanesi, tutti affiliati all'Istituto di Tecnologie Biomediche del CNR di Milano. Il processo di integrazione dei dati è centrale in Bioinformatica per la enorme quantità di informazioni presenti nei database biologici la cui interpretazione può diventare ardua se non affrontata con strumenti adeguati. In questo contesto una piattaforma Grid può essere lo strumento essenziale che permette di effettuare un processo di integrazione e un sistema di gestione dei dati ad alta efficienza necessari per completare studi che tengano in conto delle informazioni presenti su più database biologici. Questo studio tratta proprio della definizione di uno strumento innovativo per la gestione e l'integrazione di database biologici, nello specifico di due importanti database quali UNIPROT [10] e ENSEMBL [11] nella loro versione BIOMART [12], in un contesto Grid. Il cuore del progetto utilizza un Web Service che consente di effettuare contemporaneamente delle query tramite SQL (Structured Query Language) su più database distribuiti su macchine distinte utilizzando il protocollo SOAP (Simple Object Access Protocol). Questa configurazione permette all'utente connesso tramite un nodo Grid di interrogare i database e integrare i dati tramite il Web Service che rimanda l'informazione richiesta ottimizzando i tempi di comunicazione.

A livello metodologico è da sottolineare come l'idea di utilizzare la tecnologia dei Web Service per l'interrogazione dei database biologici su una piattaforma Grid è stata suggerita dal fatto che al momento non esistono strumenti per effettuare interrogazioni SQL su più database distribuiti senza interferire con l'ambiente Grid. Inoltre la necessità di utilizzare un client operante su un nodo Grid, in cui solo un numero limitato di librerie standard sono disponibili, ha spinto gli ideatori di questo strumento ad utilizzare la tecnologia Java e SOAP che garantiscono requisiti minimi di librerie addizionali e una portabilità veramente effettiva.

Un altro interessante lavoro è stato presentato alla conferenza dal gruppo del Dipartimento di Biologia dell'Università di Tor Vergata a Roma, guidato dalla Prof. Manuela Helmer-Citterich [13] con il titolo "High-throughput exploration of functional residues in protein structures". Questo lavoro si pone nel panorama degli strumenti per la ricerca delle somiglianze tra le strutture di differenti proteine, volendo associare ad una similitudine geometrica anche una relazione ed una analogia a livello funzionale tra le proteine considerate. L'innovazione e anche la chiave di questo progetto risiedono nell'elevato grado di integrazione tra i database funzionali utilizzati per i confronti, così come tra i diversi metodi per il confronto strutturale e le risorse di annotazione funzionale. All'interno del progetto è stato sviluppato un metodo per l'integrazione di tutti i maggiori database di annotazione funzionale 3D unitamente ad un algoritmo ottimizzato per la ricerca di similitudini strutturali inter-proteina. Infine i risultati ottenuti vengono catalogati in un database secondo caratteristiche chimico fisiche a livello del singolo residuo proteico e non della proteina vista come un tutto. Questo tool, chiamato PDBFUN (Fig. 3) è disponibile su server dell'Università di Roma [14] che permette un accesso al famoso database strutturale PDB (Protein Data Bank) [15] ma organizzato come database di residui annotati. Questo strumento permette, ad esempio, di identificare casi di convergenza evolutiva in strutture proteiche.

pdbfun

Figura 3 - Logo del tool pdbfun.

Sempre nel contesto dell'analisi strutturale delle proteine, uno strumento innovativo è stato senz'altro presentato al congresso da Marco Punta e Burkhard Rost, della Columbia University, con il titolo "PROFcon: prediction of internal protein contacts" (Fig. 4) [16]. Questo lavoro è stato motivato dalla crescente differenza tra il numero delle sequenze proteiche decodificate e quello delle strutture

proteiche identificate; nuovi e più efficienti strumenti per la predizione della struttura di sequenze proteiche sono dunque necessari per colmare questo divario. Il lavoro presentato mira a fornire uno strumento per migliorare il riconoscimento di contatti non-locali tra residui in una stessa proteina. Il metodo sviluppato in PROFcon per determinare i contatti intra-proteici si basa su una rete neurale che sintetizza le informazioni provenienti da varie risorse. Per la fase di addestramento e testing della rete neurale è stata usato un insieme di 3201 proteine di nota struttura estratto dalla release del Dicembre 2003 di EVA (Evaluation of protein structure prediction servers) [17]. I risultati ottenibili con questo strumento sono ovviamente legati alle caratteristiche della proteina studiata (lunghezza, classe strutturale, famiglia etc.). E' evidente che proteine più corte e con un grande numero di sequenze omologhe disponibili ottengano risultati più soddisfacenti come pure proteine che hanno nella struttura una netta maggioranza di motivi alfa-elica o beta-foglietto.

Figura 4 - Immagine logo della main page di PROFcon.

Conclusioni

La conferenza ha costituito un punto di incontro e di valutazione dello stato dell'arte della Bioinformatica a livello nazionale e la relazione con la situazione internazionale.

L'eterogenea estrazione culturale degli oratori ha evidenziato il carattere inter-disciplinare di questa scienza emergente che nasce proprio per riunire in una attività sinergica le competenze provenienti da differenti campi applicativi e teorici.

La numerosità dei partecipanti, più di un centinaio, e la partecipazione attiva dell'uditorio dimostrano quanto e quale sia l'attuale fermento in Italia per le applicazioni bioinformatiche.

Il CILEA è parte attiva di un progetto quinquennale FIRB (Fondo per gli Investimenti della Ricerca di Base) nel campo della bioinformatica. Il progetto denominato LITBIO (Laboratorio Interdisciplinare di Tecnologie Bioinformatiche) [18], avrà sede proprio all'interno della struttura del CILEA.

Bibliografia e riferimenti

[1] Pagina ufficiale della conferenza, URL: http://www.itb.cnr.it/bits2005/
[2] BITS, URL: http://www.bioinformatics.it/
[3] SIBBM, URL: http://www.fisv.org/sibbm/
[4] ABCD, URL: http://www.fisv.org/abcd/
[5] Institute for Information Trasmission Problems, Ras, Russia, URL: http://www.iitp.ru/index.html
[6] NCBI, URL: http://www.ncbi.nlm.nih.gov/
[7] Google, URL: http://google.com
[8] Scholar-Google, URL:http://scholar.google.com
[9] Progetto europeo Biosapiens, URL: http://www.biosapiens.info
[10] UNIPROT, URL: http://www.expasy.uniprot.org/
[11] ENSEMBL, URL: http://www.ensembl.org/
[12] BIOMART, URL: http://www.ebi.ac.uk/biomart/
[13] Gruppo del Dipartimento di Biologia dell'Università di Tor Vergata, URL: http://cbm.bio.uniroma2.it/
[14] Pdbfun, URL: http://pdbfun.uniroma2.it
[15] PDB, URL: http://www.pdb.org
[16] PROFcon, URL: http://www.predictprotein.org/
[17] Koh et al., Nucleic Acids Research, 31: 3311-3315, (2003)
[18] News progetto LITBIO, URL: http://www.cnr.it/cnr/news/CnrNews?IDn=1324